



DOE Systems Biology Knowledgebase

KBASE

Data and modeling for
predictive biology

Present and Future Computing Requirements

Tom Brettin and Shane Canon

Oak Ridge National Laboratory and
Lawrence Berkeley National Laboratory

NERSC BER Requirements for 2017

September 11-12, 2012

Rockville, MD



U.S. DEPARTMENT OF
ENERGY

Office of
Science

Key Personnel



Adam Arkin
Lawrence Berkeley National Lab
PI and Science



Rick Stevens
Argonne National Lab
Infrastructure (Hardware and Software)

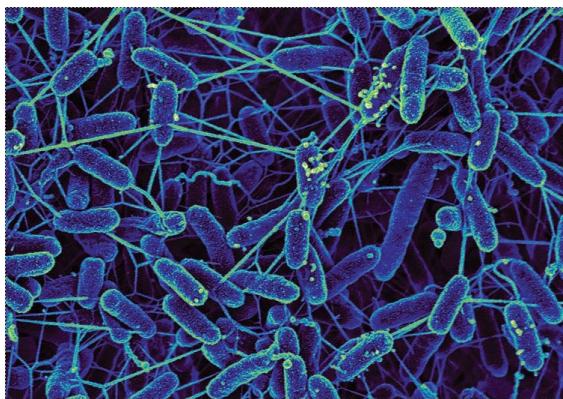


Bob Cottingham
Oak Ridge National Lab
Operations

Systems Biology Knowledge Base

Knowledgebase enabling ***predictive*** systems biology.

- Powerful modeling framework.
- **Community-driven**, extensible and scalable **open-source** software and application system.
- Infrastructure for integration and reconciliation of algorithms and data sources.
- Framework for standardization, search, and association of data.
- Enable model based **experimental design** and **interpretation** of results.



Microbes

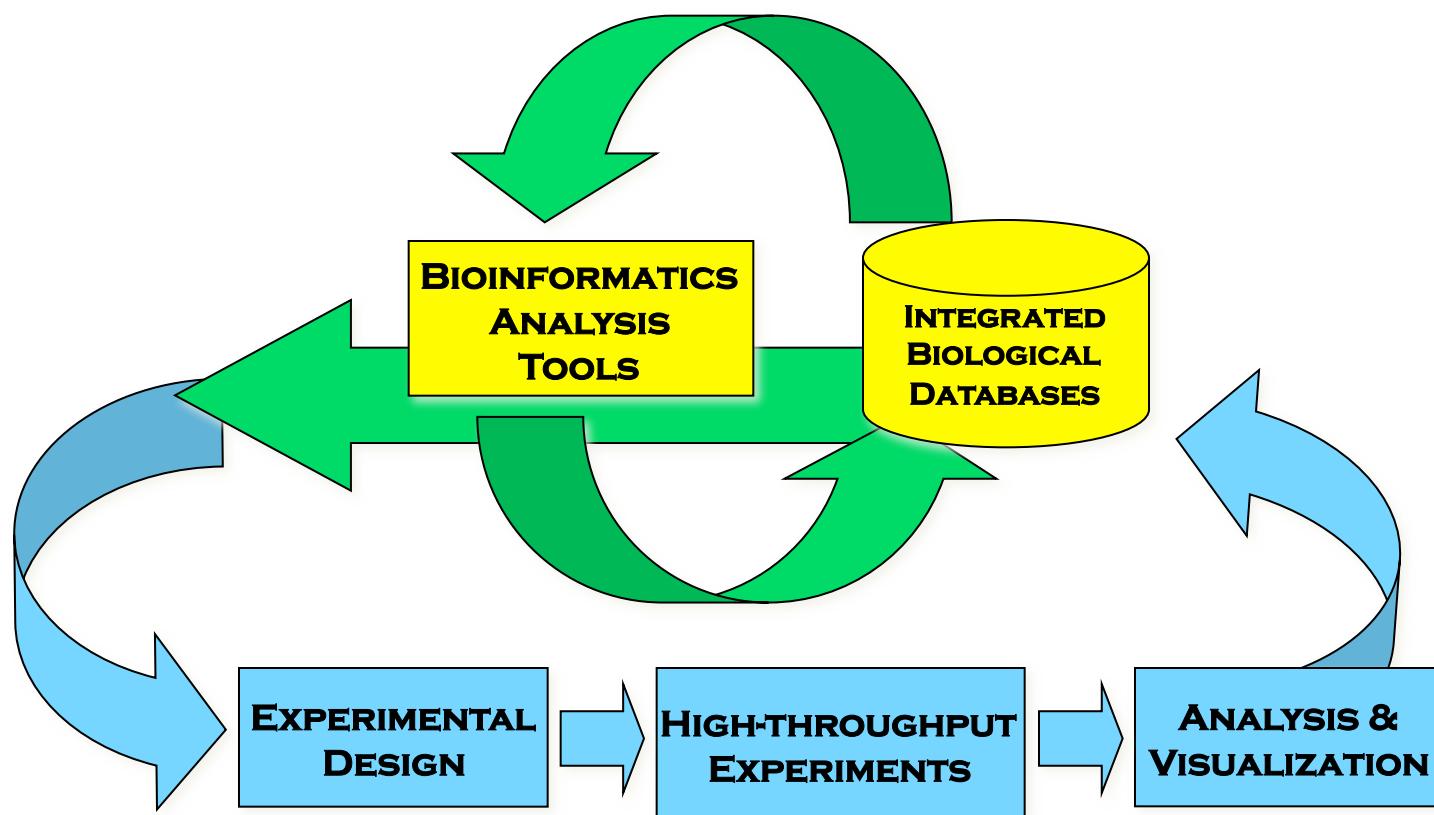


Communities

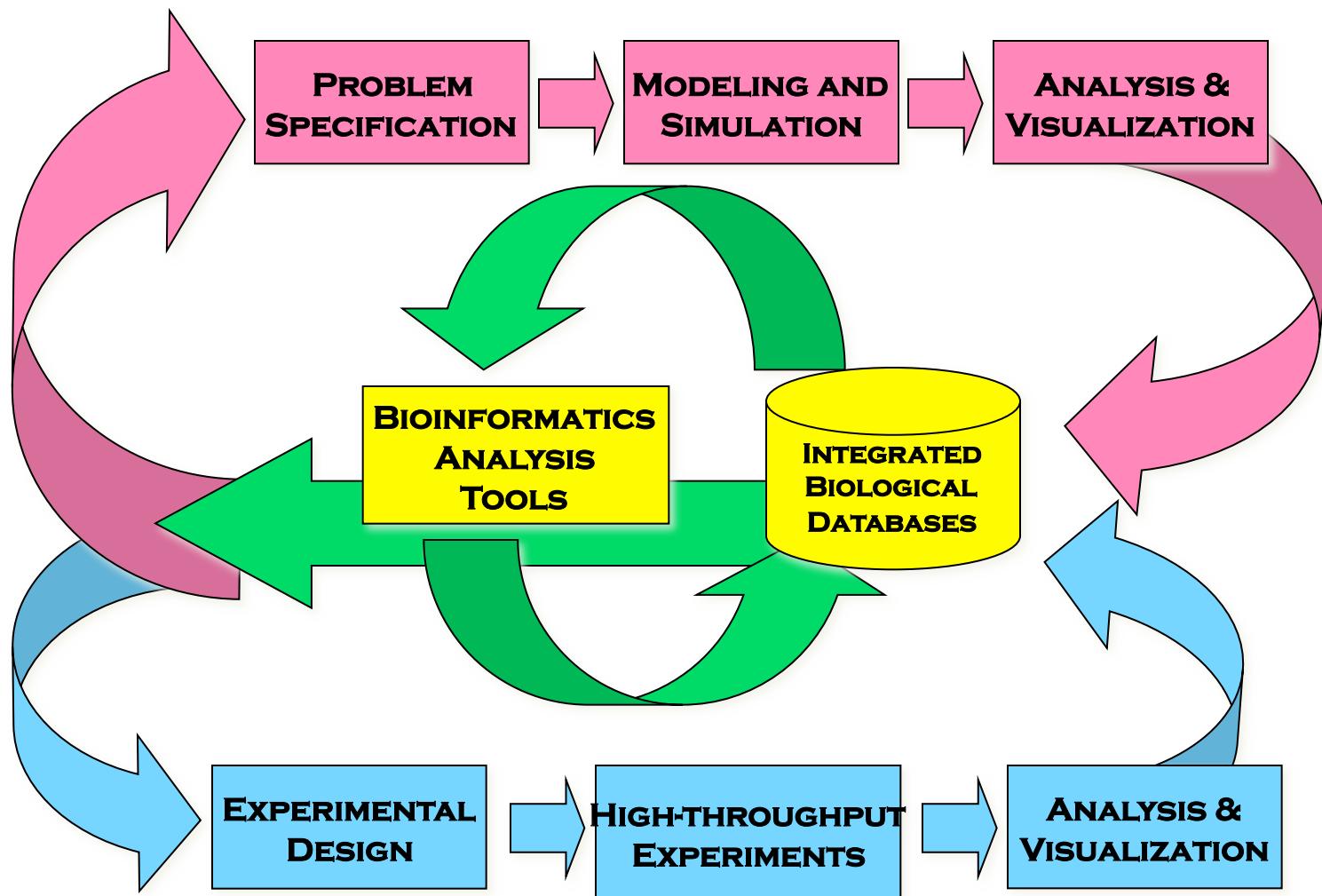


Plants

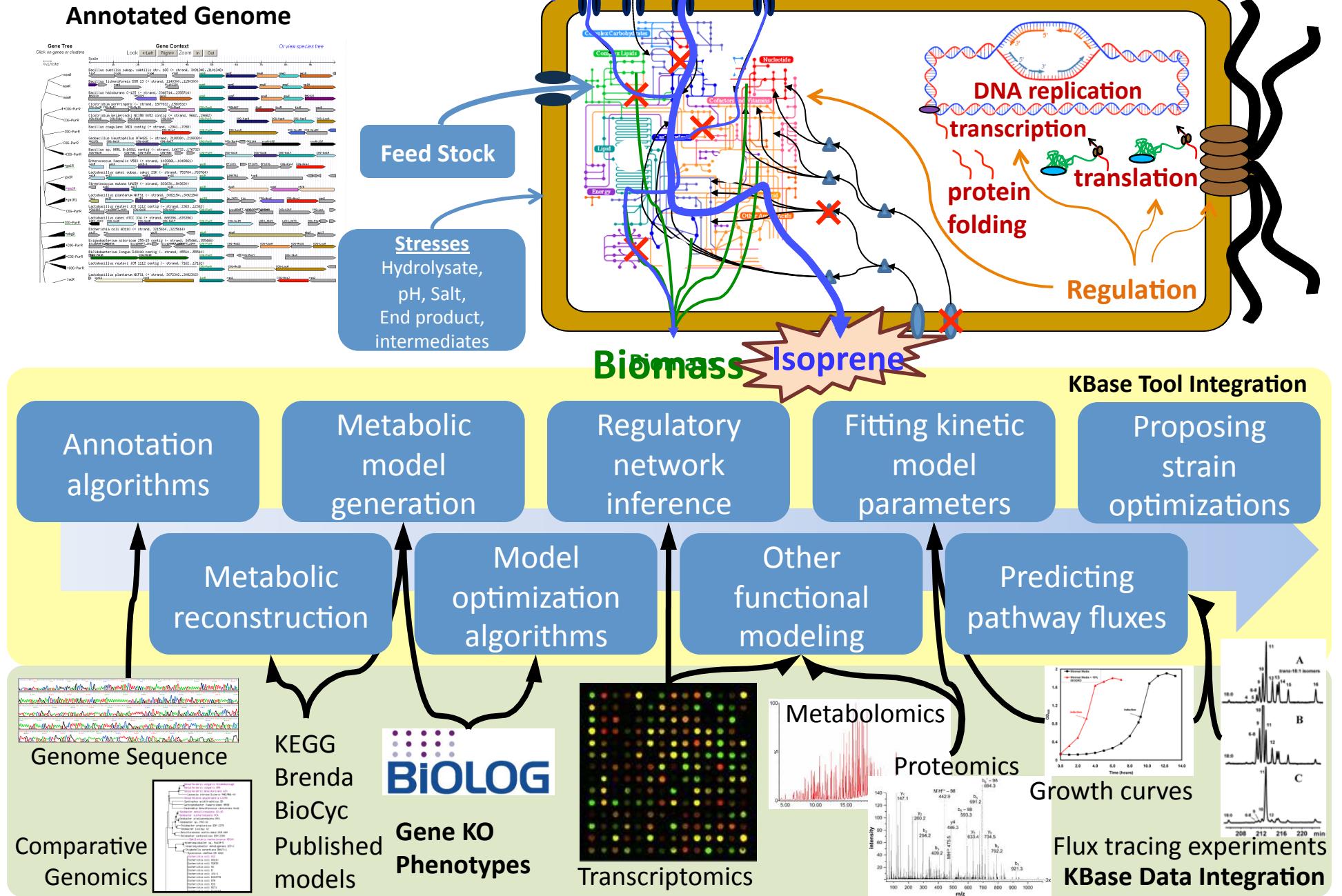
Integrated View of Modeling, Simulation, Experiment, and Bioinformatics



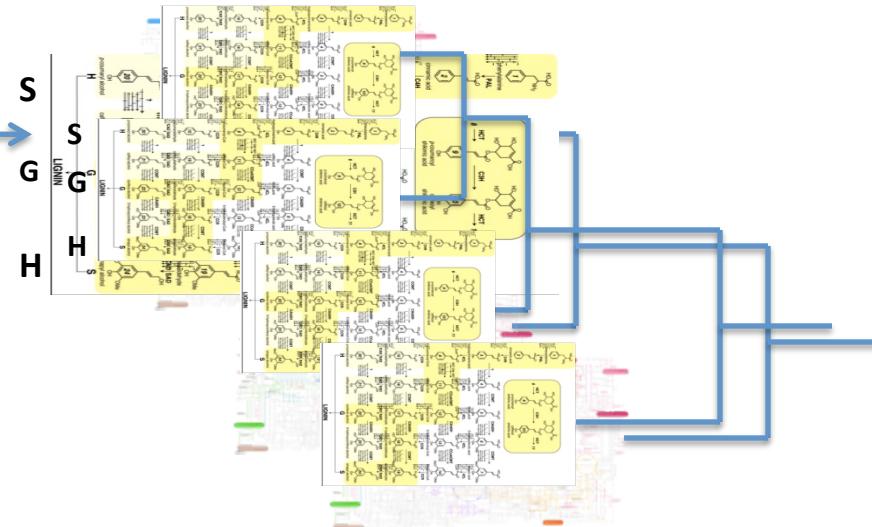
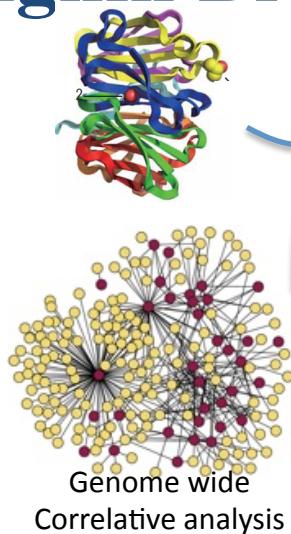
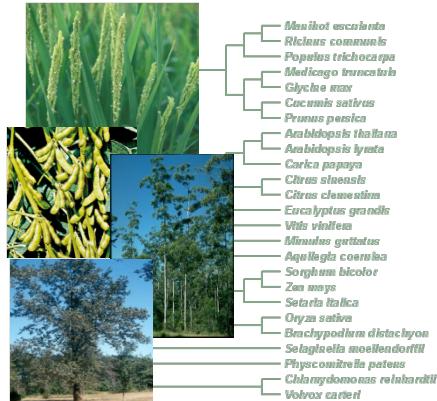
Integrated View of Modeling, Simulation, Experiment, and Bioinformatics



Engineering a Microbe for Biofuel Production



Modifying Lignin Biosynthesis



PolyPhen-2

SNP influenced changes in protein structure and function

Pathway predictions

- Model optimization
- validation

Plant systems modification

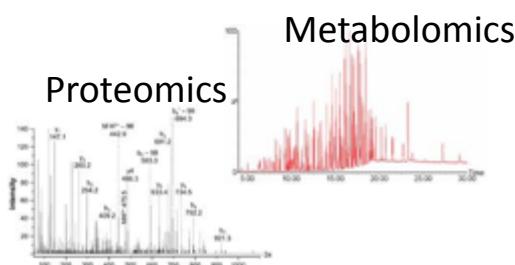
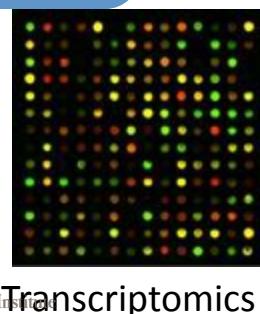
- Genome annotation algorithms
- Comparative genomics

- Network inference
- Pathway reconstruction
- Omics & SNP overlay

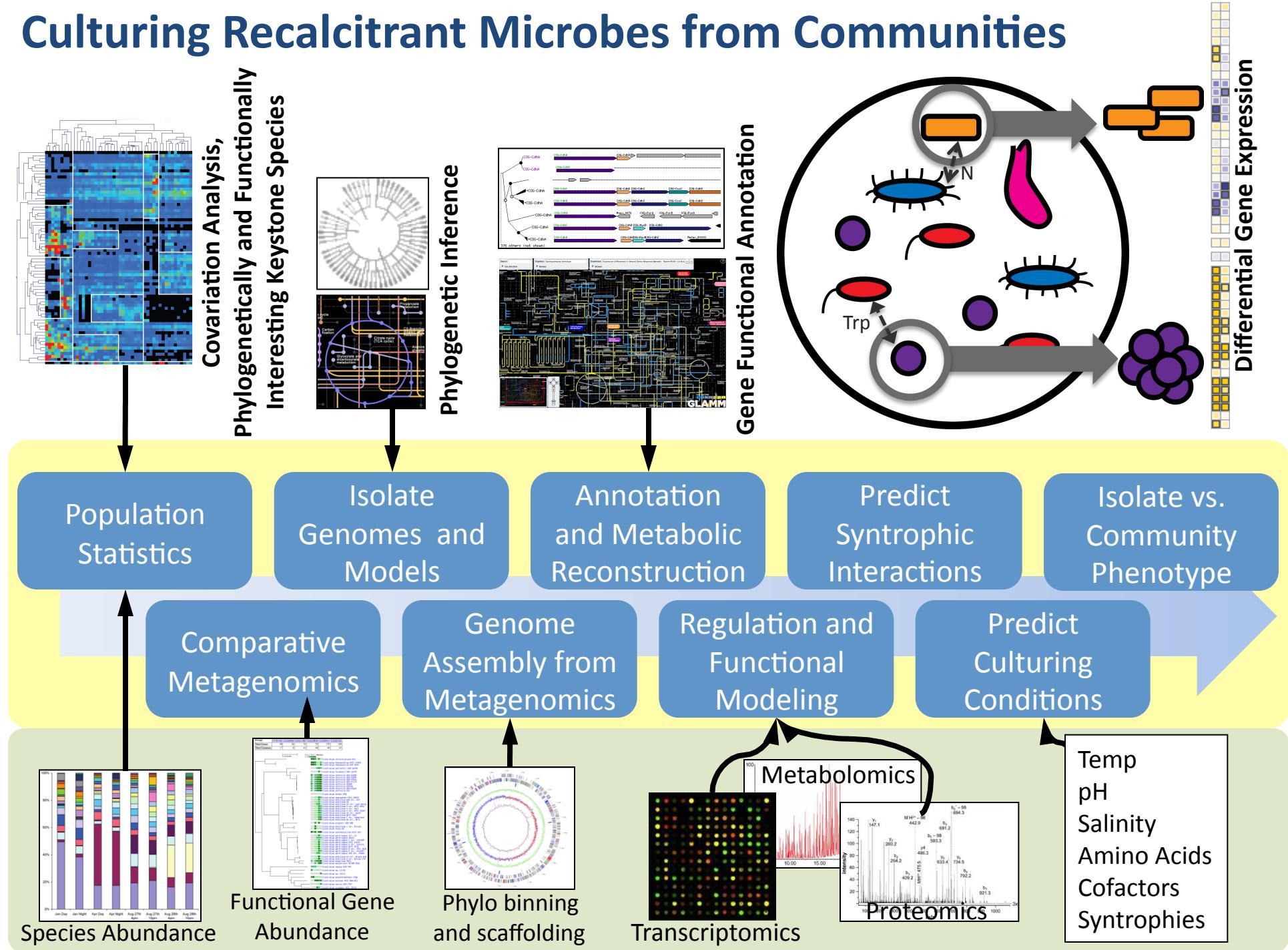
Phylogenomics
Modeling phase I



Phenotype
Mutant population
Resequencing data



Culturing Recalcitrant Microbes from Communities



Scientific Goals 2017

- Analyze understudied microbial phyla
- Interpret metagenomics data to identify conditions required for growth
- Construct, simulate and validate plant life models
- Integrate Descriptions and Annotations of Microbial Genomic Features
- Improve Plant Genome Annotation Datasets and Make Them More Accessible

Computational Strategies

Most KBase scientific targets are data driven and for the foreseeable future, would allow their computing requirements to be met by modest-scale commodity clusters.

HOWEVER, several sub problems will require substantially more computing

- Combinatorial Analysis and Optimization of Biological Networks
- Metagenome Indexing, Assembly, and Analysis
- Computing Sequence Similarities and Indexing Kbase Reference Databases

These problems are characterized by these features:

- Matrix operations
- Large aggregate memory and tightly coupled processors
- Hardware acceleration for local alignments

These problems could benefit from:

- Low level support for many-task parallelism at OS/hardware level
- Low level support for fast string comparisons and associative memory type operations
- Hardware acceleration for local alignments, K-mer indexing and associative arrays

Current NERSC HPC Usage

- To date NERSC has only been used for development and early testing
- System will be exposed via KBase Cluster Service
- Both Hopper and Carver
- Applications ported to date:
 - Sequence - aligners BLAST, BLAT
 - Assemblers - Kiki
 - Phylogenetic Clustering - RaiPHY
- In Progress
 - Metabolic Modeling - Gap Analysis
 - Assembler - Bowtie

HPC Requirements for 2017

- Very difficult to predict today
 - KBase not in general release yet
 - Needs will heavily depend on demand from community
 - Dependent on how KBase infrastructure grows
 - Dependent on level of contribution from ALCF and OLCF
- Expect similar demands to JGI once KBase is generally available?

Microbes Online

The Microbes Online homepage features a search bar for 'Sequence Search' and 'Advanced Search'. It includes sections for 'Genome actions', 'Favorites', and 'About MicrobesOnline'. A sidebar highlights 'GLAMM' (Interactive Viewer for Metabolic Pathways and Databases) and 'MicrobesOnline highlights' (3398 genomes, 1472 bacteria, 81 archaea, 163 eukaryotes).

Model SEED

The Model SEED homepage includes a search bar for 'Enter name or keyword' and a 'Find Genes' button. It displays a table of selected models (e.g., Escherichia coli K12, Sead224308) with columns for Model ID, Organism, Version, Source, Class, Genome Size, Model Size, Reactions with genes, Gapfilling with hits, and Gapfilling with hits. A 'Click here to run FBA on selected models' link is also present.

Meta Microbes Online

The Meta Microbes Online homepage features a search bar for 'Sequence Search' and 'Advanced Search'. It includes sections for 'ADD GENOMES', 'GENOMES SELECTED', 'SEARCH GENES', 'FAVORITES', and 'GENOME ACTIONS'. A sidebar highlights 'metaMicrobesOnline highlights' (3821 genomes, 1425 bacteria, 80 archaea, 120 eukaryotes, 163 metagenomes).

RegFam

The RegFam homepage includes a search bar for 'Search regulator' and a navigation menu with links to 'Regulon collections', 'Browse', 'Contact', and 'Help'. A sidebar highlights 'Regulon Collections' and 'Regulons are combined into collections of three types characterized by a common:

- I. Taxonomic group - Collection of various regulons analyzed within a particular taxonomic group of organisms.
- II. Transcription factor - Collection of cognate regulons operated by orthologous TFs in multiple taxonomic groups of genera.
- III. Metabolic pathway or biological process - Collection of multiple regulons controlling transcription of genes from a particular metabolic pathway or a biological process.

Below is a table of 'Collections of regulons by Taxonomic groups'.

	Genomes	Regulons	Transcription factors	Binding sites
Shewanella	16	82	6621	
Staphylococcus	7	47	47	1966
Streptococcus	8	39	39	1489
Thermotogae	11	31	31	626
Bacillus	11	108	108	2509
Deinococcus	10	40	40	1079
Enterobacteriales	32	58	58	16527
Cyanobacteria (draft)	14	10	10	659
Halophiles (draft)	6	15	15	514

The SEED

The SEED Viewer homepage shows an 'Organism Overview for Escherichia coli K12 (83333.1)'. It includes a table of genome statistics (TAXONOMY, SIZE, NUMBER OF CONTIGS, NUMBER OF SUBSYSTEMS, NUMBER OF CODING SEQUENCES, NUMBER OF RNAs) and a 'For each genome we offer a wide set of information to browse, compare and download' section. Below is a 'Subsystem Information' section with 'Subsystem Statistics' and 'Subsystem Coverage' charts.

MG-RAST

The MG-RAST homepage includes a search bar for 'Enter name or keyword' and a 'Find Genes' button. It displays a table of metagenomes with columns for project, biome, location, depth, country, and sequencing method. A sidebar highlights 'All Metagenomes' and 'Project (86)'.

Phytozome

The Phytozome homepage features a phylogenetic tree of plant families. A large yellow starburst contains the text '20,000+ users'. A sidebar highlights 'Species in Phytozome v7.0' and 'Announcements' (Phytozome v7.0 released, Eucalyptus grandis, Citrus sinensis (sweet orange), and clementine genomes). A news section discusses the release of version 8.0.

RAST

The RAST homepage includes a progress bar for 'Progress bar color key' and a table of 'Jobs you have access to'. The table lists jobs with columns for Job ID, User, Status, Progress, Progress Date, Annotation Progress, and Run ID.

Strategies for New Architectures

- No project plans to develop or port applications to GPU or Many Core
- Leverage community developed codes for new architectures
 - Several initiatives for GPU (especially nVidia)
 - Intel is starting to engage community
- Recent workshop on next generation sequence analysis libraries
 - Attended by vendors
 - KBase and the JGI staff represented
 - Others from the community with sequence analysis libraries and HPC experience

What new science results might be afforded by improvements in NERSC computing hardware, software and services?

- **Engineering a Microbe for Biofuel Production**
 - Ability to grow novel microorganisms in the laboratory
 - Directed engineering
- **Modifying Lignin Biosynthesis**
 - Understanding of the genetic determinants of complex phenotypic traits in eukaryotic organisms and modification for industrial and other applications
- **Culturing Recalcitrant Microbes from Communities**
 - Understanding of population structure and interactions in complex mixtures of microorganisms and replicating these populations in laboratory settings.



DOE Systems Biology Knowledgebase

Thank You

Additional Reading:

DOE Systems Biology Knowledge Base
Implementation Plan